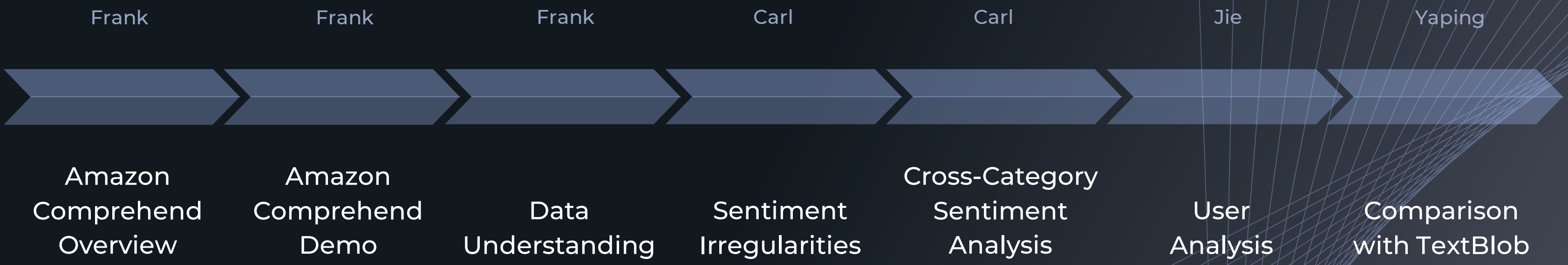
EMORY GOIZUETA BUSINESS SCHOOL | Master of Science in Business Analytics MSBA

# Managing Big Data : Amazon Comprehend & NLP with Amazon Reviews

amazon 404

# Presentation Overview



Frank

Amazon Comprehend Overview

Frank

Amazon Comprehend Demo

Frank

Data Understanding

Carl

Sentiment Irregularities

Carl

Cross-Category Sentiment Analysis

Jie

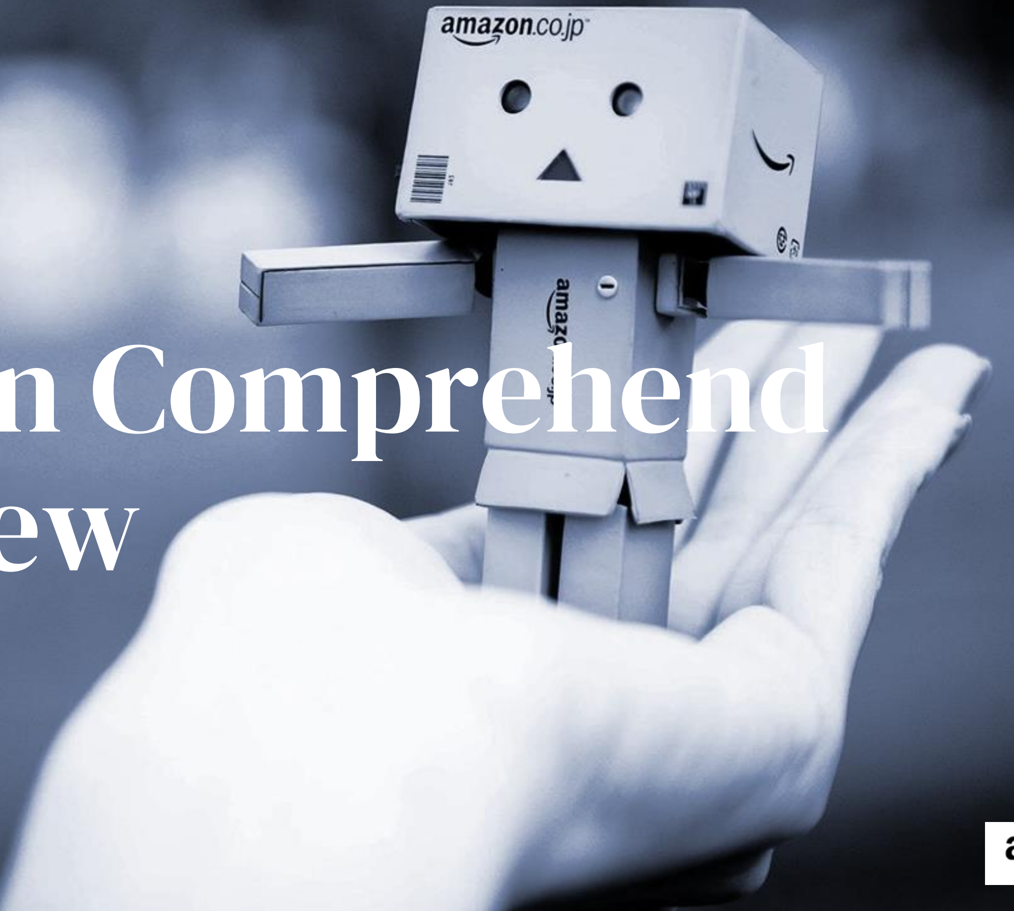
User Analysis

Yaping

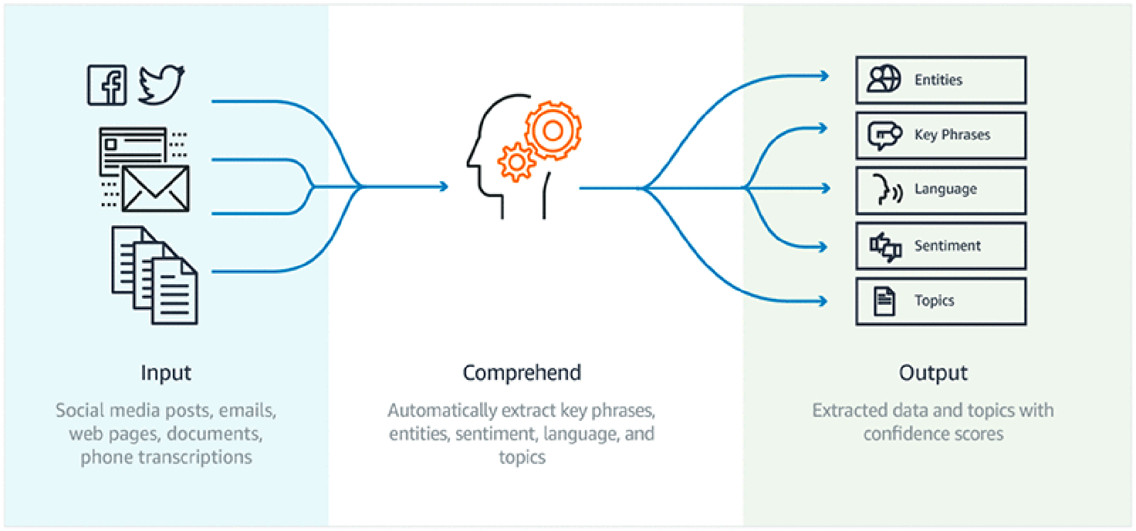
Comparison with TextBlob

# 1.

# Amazon Comprehend Overview



“ **Amazon Comprehend** is a natural language processing (NLP) service that uses machine learning to discover insights from text.



Input

Social media posts, emails, web pages, documents, phone transcriptions

Comprehend

Automatically extract key phrases, entities, sentiment, language, and topics

Output

Extracted data and topics with confidence scores

# How does it work:

Amazon Comprehend uses a pre-trained model to examine and analyze a document or set of documents to gather insights about it. This model is continuously trained on a large body of text so that there is no need for you to provide training data.

# Features

### Keyphrase Extraction

Amazon Comprehend extracts key phrases that appear in a document. For example, a document about a basketball game might return the names of the teams, the name of the venue, and the final score.

### Entity Recognition

Amazon Comprehend returns a list of entities, such as people, places, and locations, identified in a document.

### Detect the Dominant Language

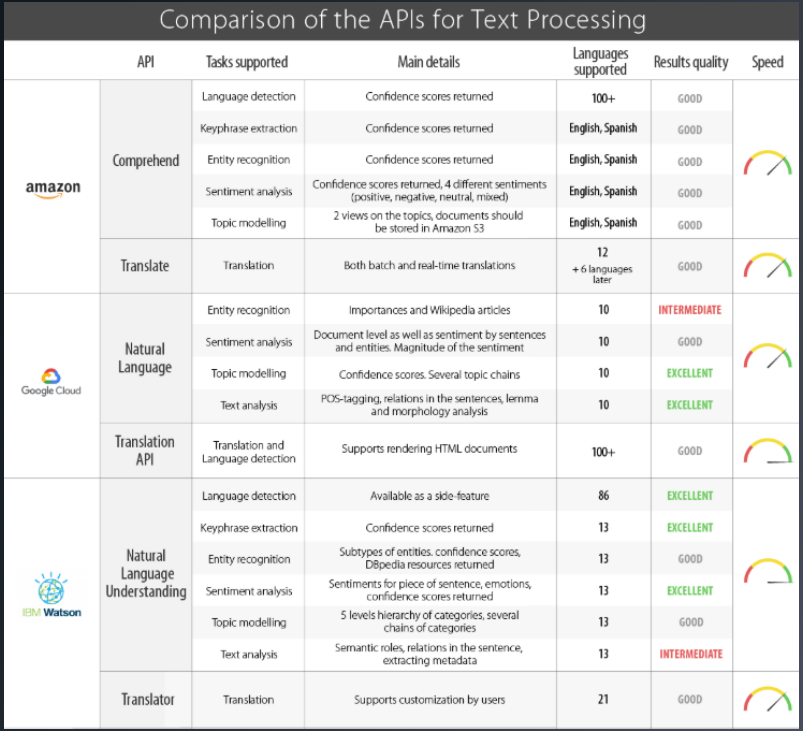
Amazon Comprehend identifies the dominant language in a document. Amazon Comprehend can identify over 100 languages.

### Sentiment Analysis

Amazon Comprehend determines the emotional sentiment of a document. Sentiment can be positive, neutral, negative, or mixed.

# Comparison of the APIs for NLP

- Amazon Comprehend
- Google Cloud Analytics
- IBM Watson Natural Language Classifier



Comparison of the APIs for Text Processing

| | API | Tasks supported | Main details | Languages supported | Results quality | Speed |
|---|---|---|---|---|---|---|
| amazon | Comprehend | Language detection | Confidence scores returned | 100+ | GOOD | |
| | | Keyphrase extraction | Confidence scores returned | English, Spanish | GOOD | |
| | | Entity recognition | Confidence scores returned | English, Spanish | GOOD | |
| | | Sentiment analysis | Confidence scores returned, 4 different sentiments (positive, negative, neutral, mixed) | English, Spanish | GOOD | |
| | | Topic modelling | 2 views on the topics, documents should be stored in Amazon S3 | English, Spanish | GOOD | |
| | Translate | Translation | Both batch and real-time translations | 12<br>+ 6 languages later | GOOD | |
| Google Cloud | Natural Language | Entity recognition | Importances and Wikipedia articles | 10 | INTERMEDIATE | |
| | | Sentiment analysis | Document level as well as sentiment by sentences and entities. Magnitude of the sentiment | 10 | GOOD | |
| | | Topic modelling | Confidence scores. Several topic chains | 10 | EXCELLENT | |
| | | Text analysis | POS-tagging, relations in the sentences, lemma and morphology analysis | 10 | EXCELLENT | |
| | Translation API | Translation and Language detection | Supports rendering HTML documents | 100+ | GOOD | |
| IBM Watson | Natural Language Understanding | Language detection | Available as a side-feature | 86 | EXCELLENT | |
| | | Keyphrase extraction | Confidence scores returned | 13 | EXCELLENT | |
| | | Entity recognition | Subtypes of entities. confidence scores, DBpedia resources returned | 13 | GOOD | |
| | | Sentiment analysis | Sentiments for piece of sentence, emotions, confidence scores returned | 13 | EXCELLENT | |
| | | Topic modelling | 5 levels hierarchy of categories, several chains of categories | 13 | GOOD | |
| | | Text analysis | Semantic roles, relations in the sentence, extracting metadata | 13 | INTERMEDIATE | |
| | Translator | Translation | Supports customization by users | 21 | GOOD | |

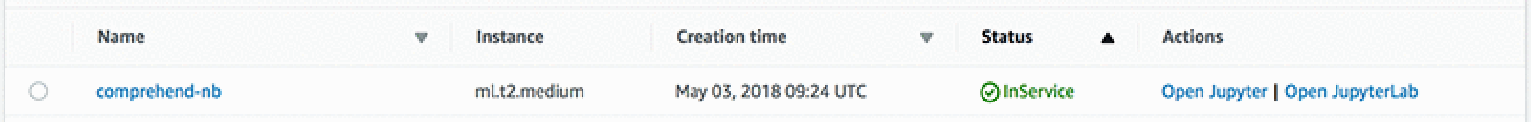
# 2.

# Amazon Comprehend Demonstration

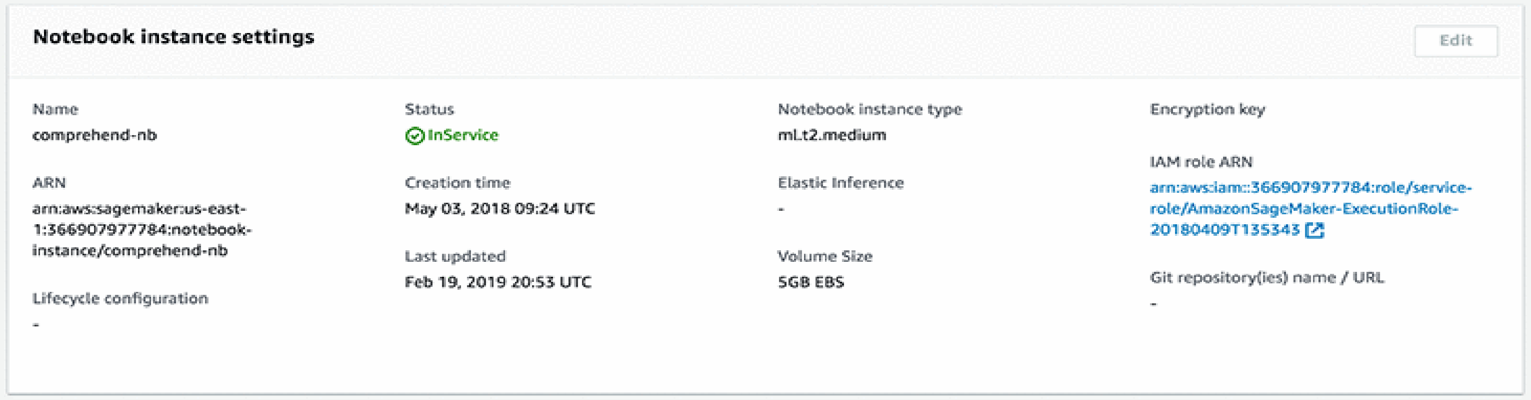
# Getting Started

### Step 1: Set up your Amazon SageMaker notebook

From the AWS Management Console, choose Services and then Amazon SageMaker under Machine Learning, and in the Amazon SageMaker console, under Notebook, choose Notebook instances. Now choose the Create Notebook Instance. And, in the console, under IAM(Identity and Access Management) role, choose Create a new role.



| | Name ▼ | Instance | Creation time ▼ | Status ▲ | Actions |
|---|---|---|---|---|---|
| ○ | comprehend-nb | ml.t2.medium | May 03, 2018 09:24 UTC | ✓ InService | Open Jupyter \| Open JupyterLab |



**Notebook instance settings**

Name
comprehend-nb

ARN
arn:aws:sagemaker:us-east-1:366907977784:notebook-instance/comprehend-nb

Lifecycle configuration
-

Status
✓ InService

Creation time
May 03, 2018 09:24 UTC

Last updated
Feb 19, 2019 20:53 UTC

Notebook instance type
ml.t2.medium

Elastic Inference
-

Volume Size
5GB EBS

Encryption key

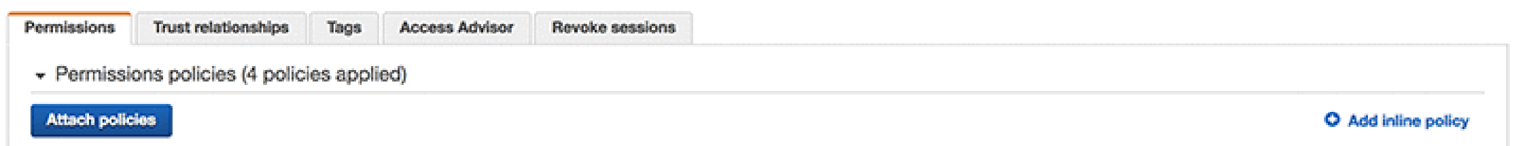
IAM role ARN
arn:aws:iam::366907977784:role/service-role/AmazonSageMaker-ExecutionRole-20180409T135343

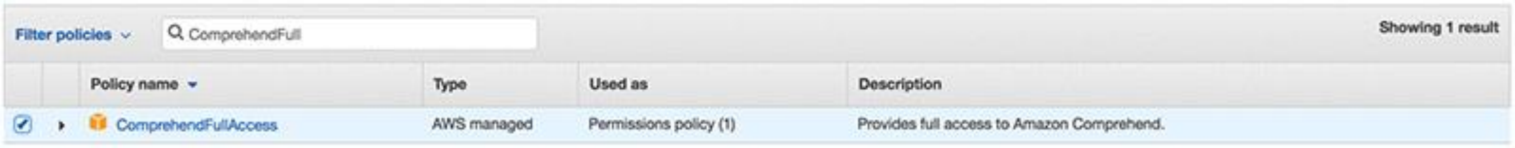
Git repository(ies) name / URL
-

# Getting Started

## Step 2: Attach policies

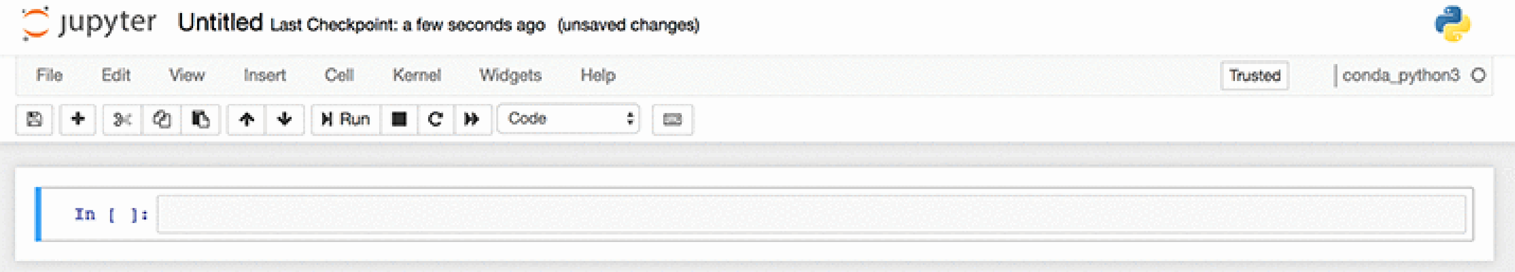
From the IAM dashboard, choose Attach policies.





| | | Policy name | Type | Used as | Description |
|---|---|---|---|---|---|
| ☑ | ▸ | ComprehendFullAccess | AWS managed | Permissions policy (1) | Provides full access to Amazon Comprehend. |

## Step 3: Create a notebook



After you open the notebook instance that you provisioned, from the Jupyter console, choose New and then conda_python3.

# Getting Started

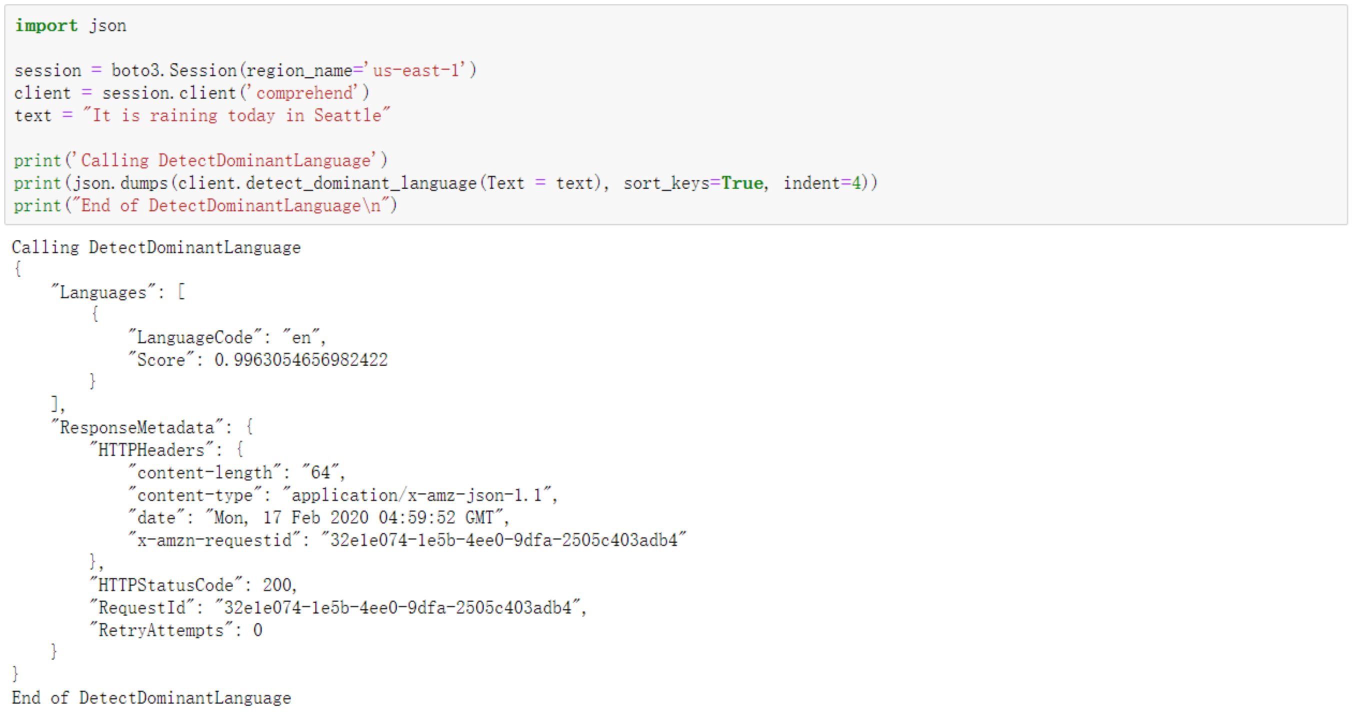
### Step 4: Connect to Amazon Comprehend

Then we can use the AWS SDK for Python SDK (Boto3) to connect to Amazon Comprehend from your Python code base. Boto is the Amazon Web Services (AWS) SDK for Python. It enables Python developers to create, configure, and manage AWS services, such as S3 and Amazon Comprehend. Using the following command, we import boto3 and connect to Amazon Comprehend in a specified AWS Region using the boto3 client.

```
import boto3

comprehend = boto3.client('comprehend',region_name='us-east-1')
```

# Detect the Dominant Language

Amazon Comprehend can automatically identifies text written in over 100 languages and returns the dominant language with a confidence score to support that a language is dominant.

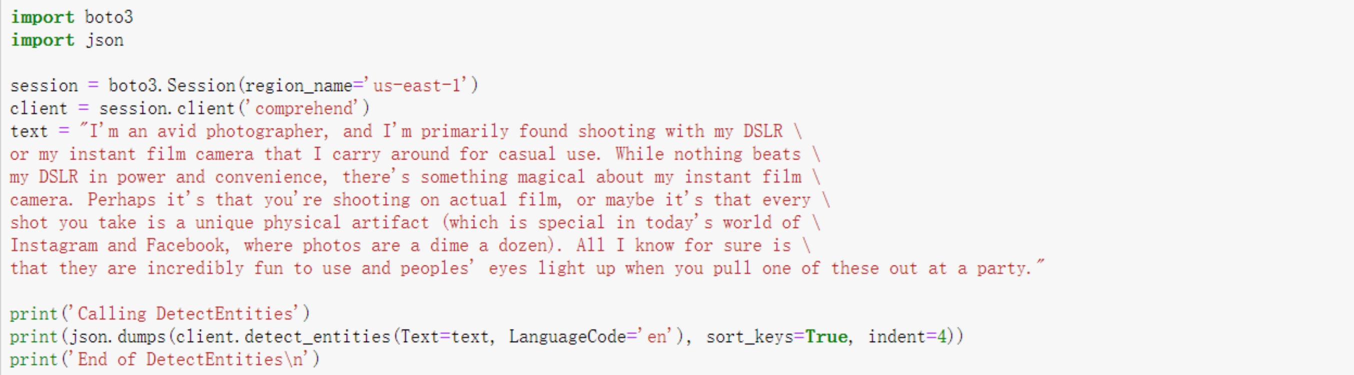


```
import json

session = boto3.Session(region_name='us-east-1')
client = session.client('comprehend')
text = "It is raining today in Seattle"

print('Calling DetectDominantLanguage')
print(json.dumps(client.detect_dominant_language(Text = text), sort_keys=True, indent=4))
print("End of DetectDominantLanguage\n")
```

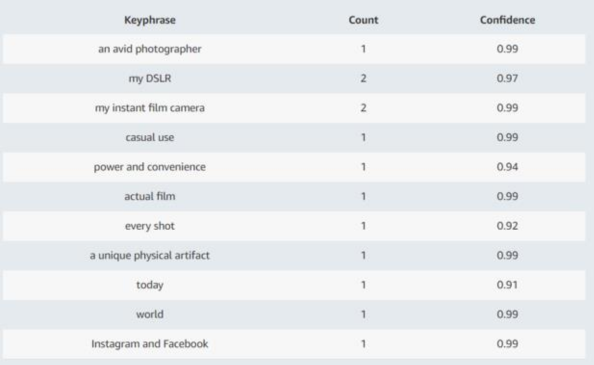
```
Calling DetectDominantLanguage
{
    "Languages": [
        {
            "LanguageCode": "en",
            "Score": 0.9963054656982422
        }
    ],
    "ResponseMetadata": {
        "HTTPHeaders": {
            "content-length": "64",
            "content-type": "application/x-amz-json-1.1",
            "date": "Mon, 17 Feb 2020 04:59:52 GMT",
            "x-amzn-requestid": "32e1e074-1e5b-4ee0-9dfa-2505c403adb4"
        },
        "HTTPStatusCode": 200,
        "RequestId": "32e1e074-1e5b-4ee0-9dfa-2505c403adb4",
        "RetryAttempts": 0
    }
}
End of DetectDominantLanguage
```

# Keyphrase Extraction

Amazon Comprehend returns the key phrases or talking points and a confidence score to support that this is a key phrase.



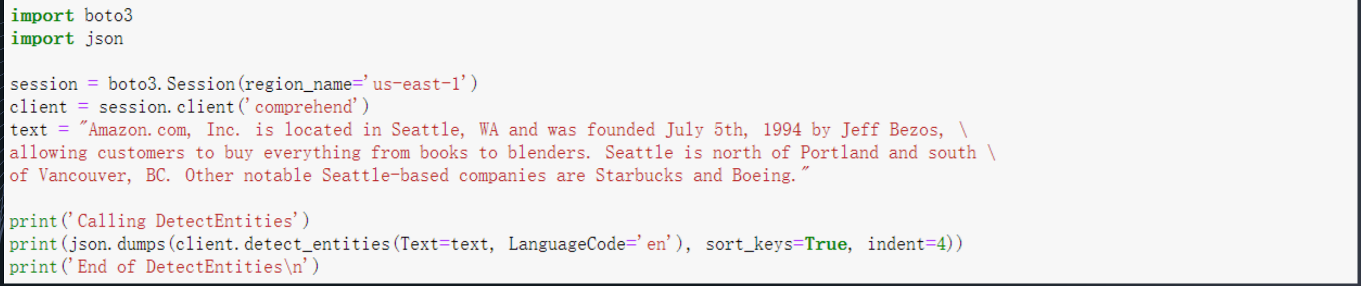
```
import boto3
import json

session = boto3.Session(region_name='us-east-1')
client = session.client('comprehend')
text = "I'm an avid photographer, and I'm primarily found shooting with my DSLR \
or my instant film camera that I carry around for casual use. While nothing beats \
my DSLR in power and convenience, there's something magical about my instant film \
camera. Perhaps it's that you're shooting on actual film, or maybe it's that every \
shot you take is a unique physical artifact (which is special in today's world of \
Instagram and Facebook, where photos are a dime a dozen). All I know for sure is \
that they are incredibly fun to use and peoples' eyes light up when you pull one of these out at a party."

print('Calling DetectEntities')
print(json.dumps(client.detect_entities(Text=text, LanguageCode='en'), sort_keys=True, indent=4))
print('End of DetectEntities\n')
```



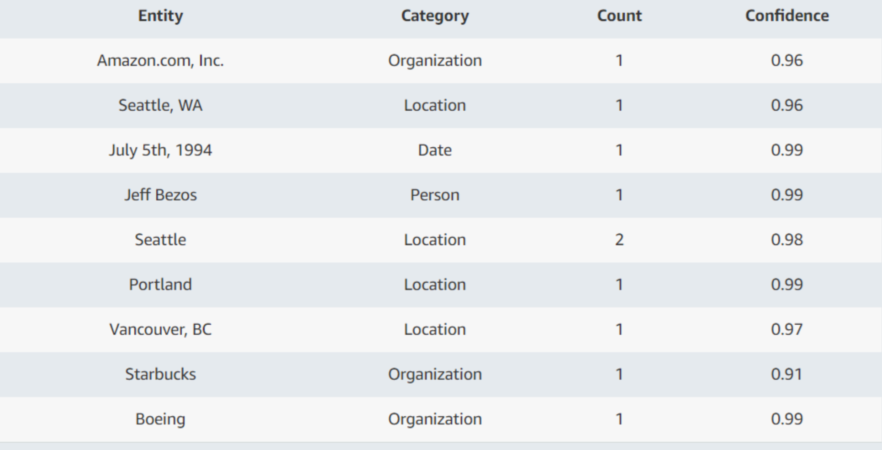
| Keyphrase | Count | Confidence |
|---|---|---|
| an avid photographer | 1 | 0.99 |
| my DSLR | 2 | 0.97 |
| my instant film camera | 2 | 0.99 |
| casual use | 1 | 0.99 |
| power and convenience | 1 | 0.94 |
| actual film | 1 | 0.99 |
| every shot | 1 | 0.92 |
| a unique physical artifact | 1 | 0.99 |
| today | 1 | 0.91 |
| world | 1 | 0.99 |
| Instagram and Facebook | 1 | 0.99 |

# Entity Recognition

The Entity Recognition API returns the named entities ("People," "Places," "Locations," etc.) that are automatically categorized based on the provided text.



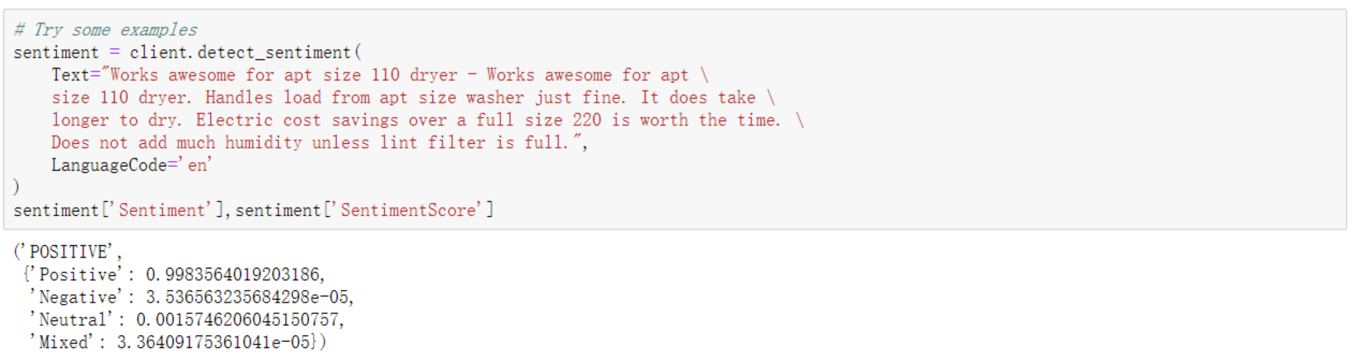
```
import boto3
import json

session = boto3.Session(region_name='us-east-1')
client = session.client('comprehend')
text = "Amazon.com, Inc. is located in Seattle, WA and was founded July 5th, 1994 by Jeff Bezos, \
allowing customers to buy everything from books to blenders. Seattle is north of Portland and south \
of Vancouver, BC. Other notable Seattle-based companies are Starbucks and Boeing."

print('Calling DetectEntities')
print(json.dumps(client.detect_entities(Text=text, LanguageCode='en'), sort_keys=True, indent=4))
print('End of DetectEntities\n')
```



| Entity | Category | Count | Confidence |
|---|---|---|---|
| Amazon.com, Inc. | Organization | 1 | 0.96 |
| Seattle, WA | Location | 1 | 0.96 |
| July 5th, 1994 | Date | 1 | 0.99 |
| Jeff Bezos | Person | 1 | 0.99 |
| Seattle | Location | 2 | 0.98 |
| Portland | Location | 1 | 0.99 |
| Vancouver, BC | Location | 1 | 0.97 |
| Starbucks | Organization | 1 | 0.91 |
| Boeing | Organization | 1 | 0.99 |

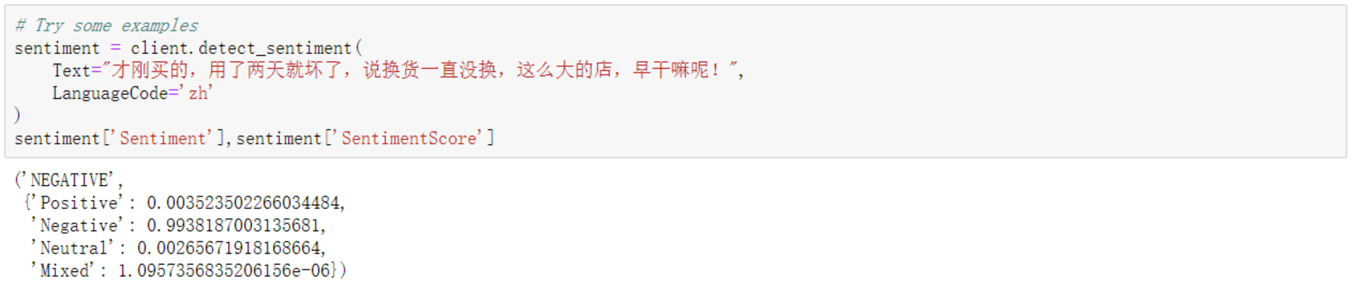
# Sentiment Analysis

The following Python program detects the sentiment of input text. You must specify the language of the input text.



```
# Try some examples
sentiment = client.detect_sentiment(
    Text="Works awesome for apt size 110 dryer - Works awesome for apt \
    size 110 dryer. Handles load from apt size washer just fine. It does take \
    longer to dry. Electric cost savings over a full size 220 is worth the time. \
    Does not add much humidity unless lint filter is full.",
    LanguageCode='en'
)
sentiment['Sentiment'],sentiment['SentimentScore']
```

```
('POSITIVE',
 {'Positive': 0.9983564019203186,
  'Negative': 3.536563235684298e-05,
  'Neutral': 0.0015746206045150757,
  'Mixed': 3.36409175361041e-05})
```

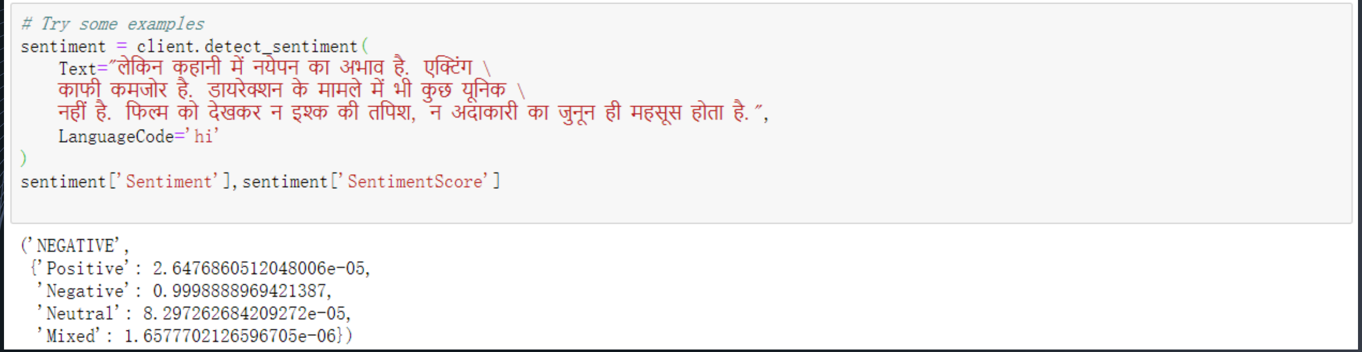


```
# Try some examples
sentiment = client.detect_sentiment(
    Text="才刚买的，用了两天就坏了，说换货一直没换，这么大的店，早干嘛呢！",
    LanguageCode='zh'
)
sentiment['Sentiment'],sentiment['SentimentScore']
```

```
('NEGATIVE',
 {'Positive': 0.003523502266034484,
  'Negative': 0.9938187003135681,
  'Neutral': 0.00265671918168664,
  'Mixed': 1.0957356835206156e-06})
```

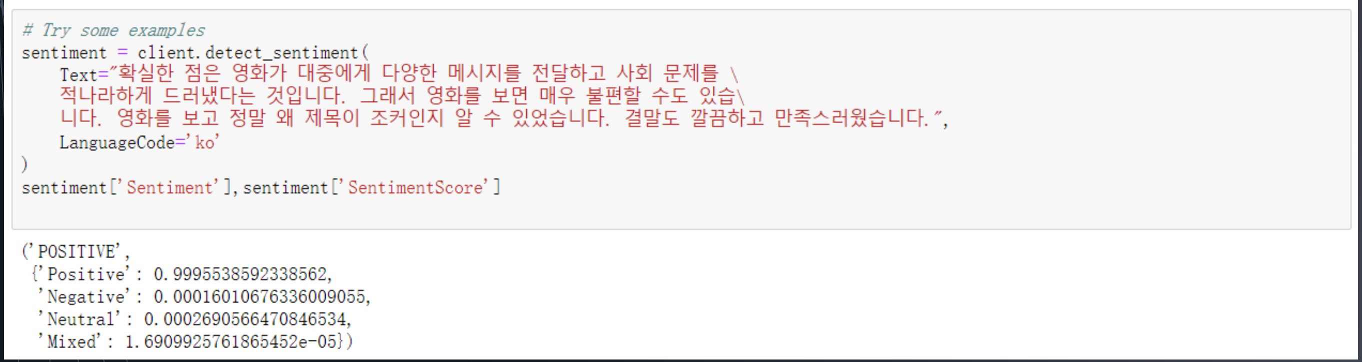
# Sentiment Analysis

The following Python program detects the sentiment of input text. You must specify the language of the input text.



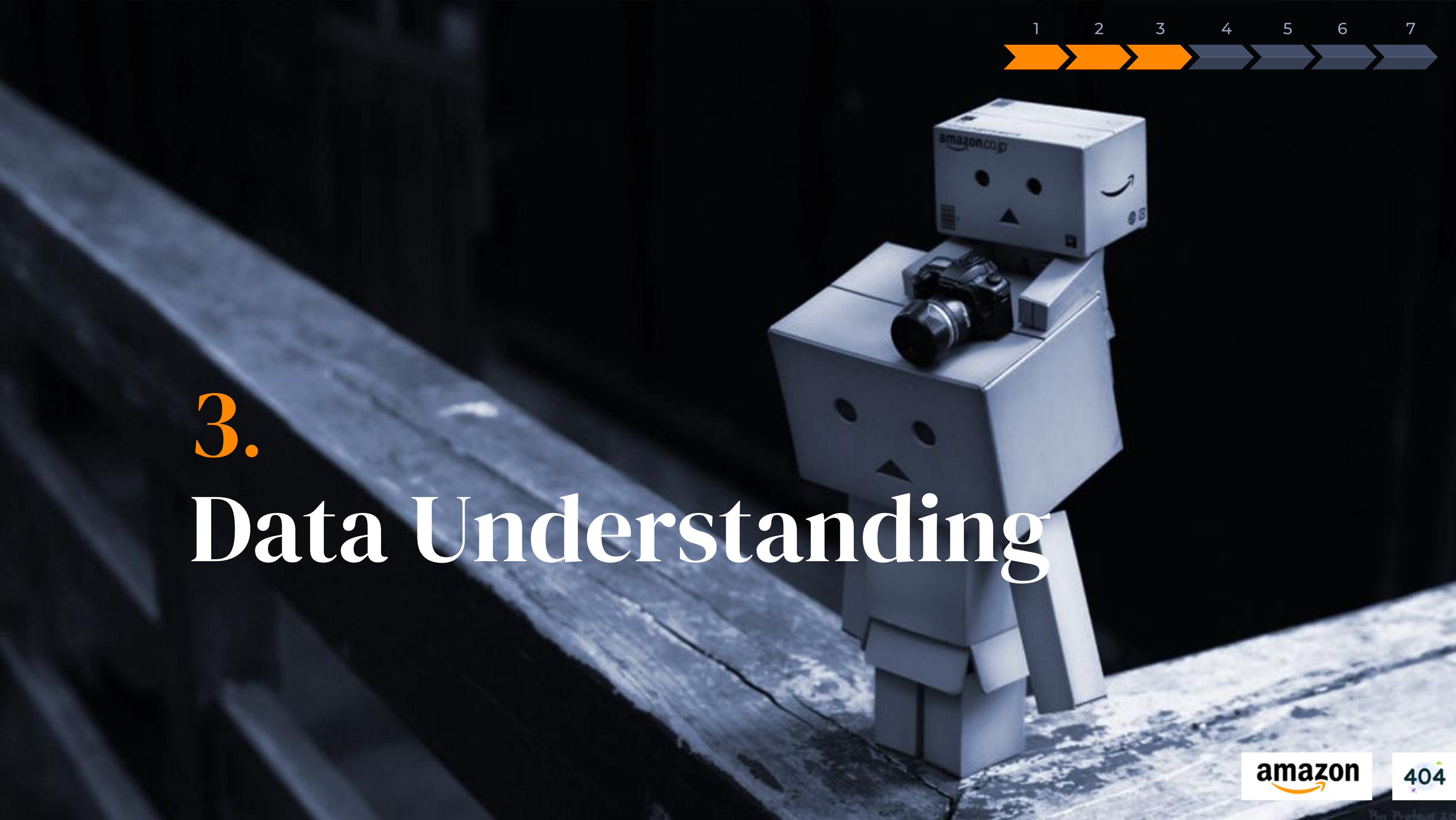
```
# Try some examples
sentiment = client.detect_sentiment(
    Text="लेकिन कहानी में नयेपन का अभाव है. एक्टिंग \
    काफी कमजोर है. डायरेक्शन के मामले में भी कुछ यूनिक \
    नहीं है. फिल्म को देखकर न इश्क की तपिश, न अदाकारी का जुनून ही महसूस होता है.",
    LanguageCode='hi'
)
sentiment['Sentiment'],sentiment['SentimentScore']
```

```
('NEGATIVE',
 {'Positive': 2.6476860512048006e-05,
  'Negative': 0.9998888969421387,
  'Neutral': 8.297262684209272e-05,
  'Mixed': 1.6577702126596705e-06})
```



```
# Try some examples
sentiment = client.detect_sentiment(
    Text="확실한 점은 영화가 대중에게 다양한 메시지를 전달하고 사회 문제를 \
    적나라하게 드러냈다는 것입니다. 그래서 영화를 보면 매우 불편할 수도 있습\
    니다. 영화를 보고 정말 왜 제목이 조커인지 알 수 있었습니다. 결말도 깔끔하고 만족스러웠습니다.",
    LanguageCode='ko'
)
sentiment['Sentiment'],sentiment['SentimentScore']
```

```
('POSITIVE',
 {'Positive': 0.9995538592338562,
  'Negative': 0.00016010676336009055,
  'Neutral': 0.0002690566470846534,
  'Mixed': 1.6909925761865452e-05})
```

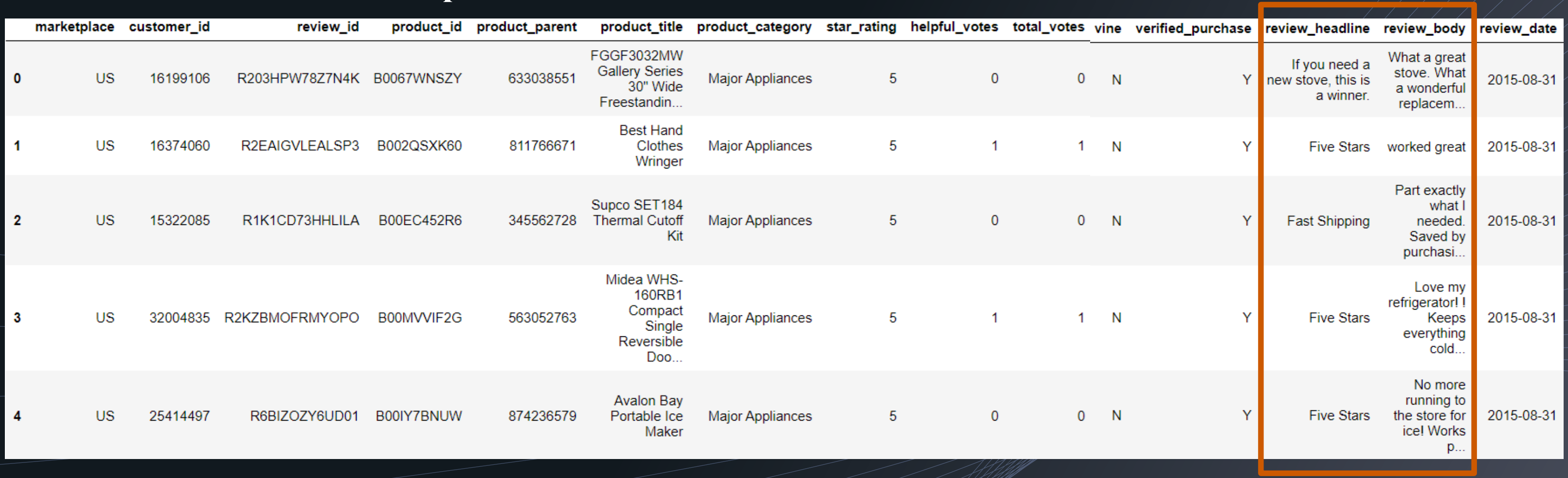


# 3.

# Data Understanding

# Amazon Customer Reviews:

**Amazon Customer Reviews (a.k.a. Product Reviews) is one of Amazon's iconic products. In a period of over two decades since the first review in 1995, millions of Amazon customers have contributed over a hundred million reviews to express opinions and describe their experiences regarding products on the Amazon.com website. Over 130+ million customer reviews are available to researchers as part of this dataset.**



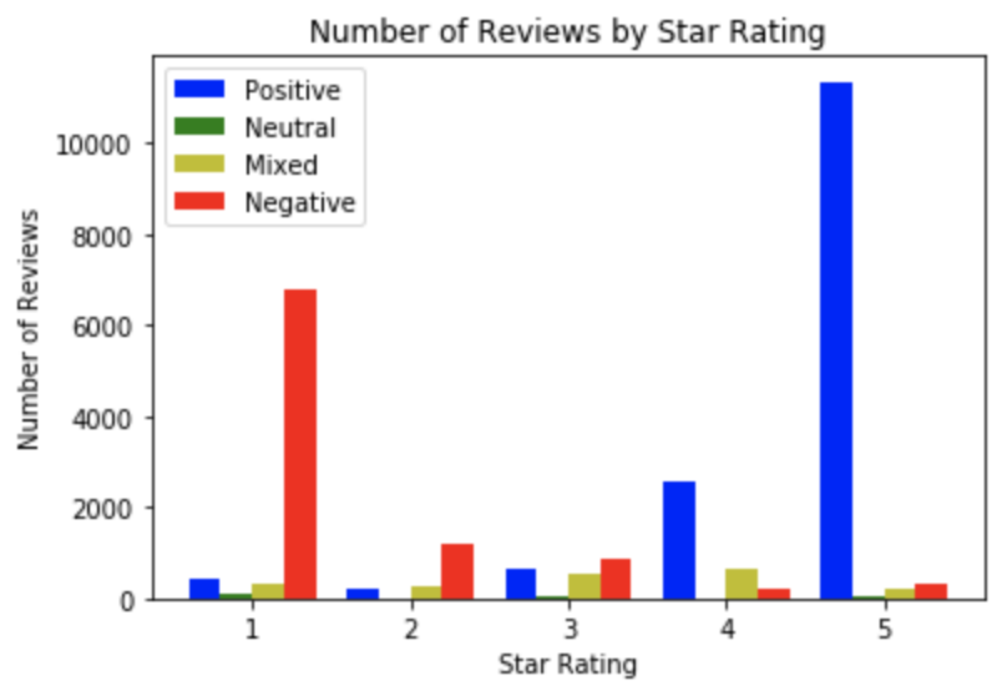
| | marketplace | customer_id | review_id | product_id | product_parent | product_title | product_category | star_rating | helpful_votes | total_votes | vine | verified_purchase | review_headline | review_body | review_date |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **0** | US | 16199106 | R203HPW78Z7N4K | B0067WNSZY | 633038551 | FGGF3032MW Gallery Series 30" Wide Freestandin... | Major Appliances | 5 | 0 | 0 | N | Y | If you need a new stove, this is a winner. | What a great stove. What a wonderful replacem... | 2015-08-31 |
| **1** | US | 16374060 | R2EAIGVLEALSP3 | B002QSXK60 | 811766671 | Best Hand Clothes Wringer | Major Appliances | 5 | 1 | 1 | N | Y | Five Stars | worked great | 2015-08-31 |
| **2** | US | 15322085 | R1K1CD73HHLILA | B00EC452R6 | 345562728 | Supco SET184 Thermal Cutoff Kit | Major Appliances | 5 | 0 | 0 | N | Y | Fast Shipping | Part exactly what I needed. Saved by purchasi... | 2015-08-31 |
| **3** | US | 32004835 | R2KZBMOFRMYOPO | B00MVVIF2G | 563052763 | Midea WHS-160RB1 Compact Single Reversible Doo... | Major Appliances | 5 | 1 | 1 | N | Y | Five Stars | Love my refrigerator! ! Keeps everything cold... | 2015-08-31 |
| **4** | US | 25414497 | R6BIZOZY6UD01 | B00IY7BNUW | 874236579 | Avalon Bay Portable Ice Maker | Major Appliances | 5 | 0 | 0 | N | Y | Five Stars | No more running to the store for ice! Works p... | 2015-08-31 |

- * Amazon Vine invites the most trusted reviewers on Amazon to post opinions about new and pre-release items to help their fellow customers make informed purchase decisions. Amazon invites customers to become Vine Voices based on their reviewer rank, which is a reflection of the quality and helpfulness of their reviews as judged by other Amazon customers.

# 4.

# Sentiment Irregularities

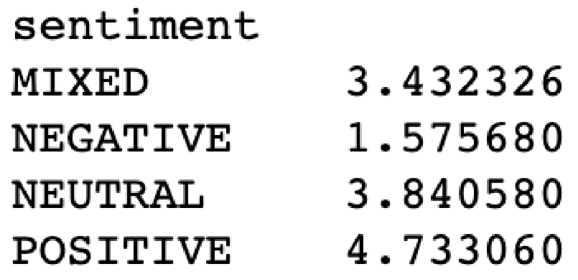
# Overall Star Rating Distribution



**Average Star Rating:**

**3.8853**

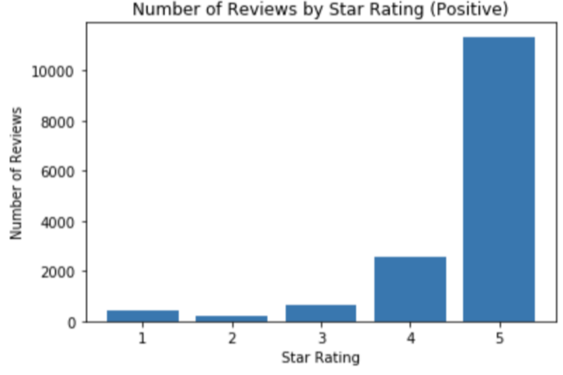
**Average Star Rating by Sentiment:**

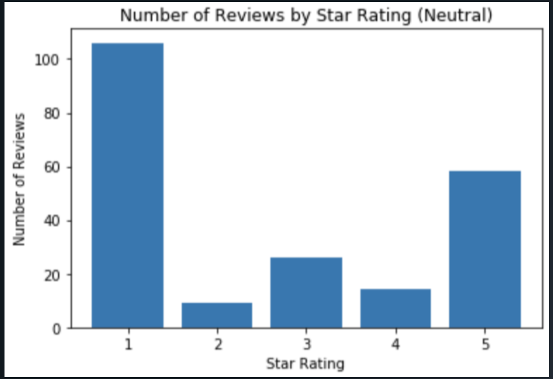


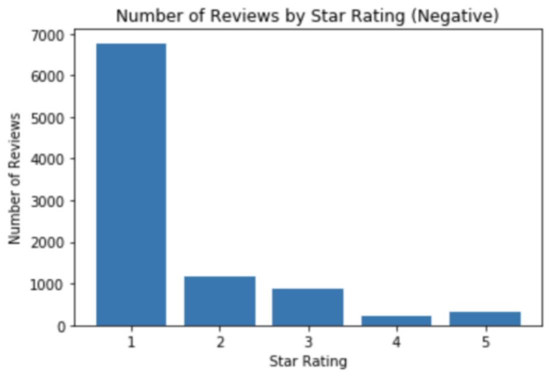
```
sentiment
MIXED        3.432326
NEGATIVE     1.575680
NEUTRAL      3.840580
POSITIVE     4.733060
```

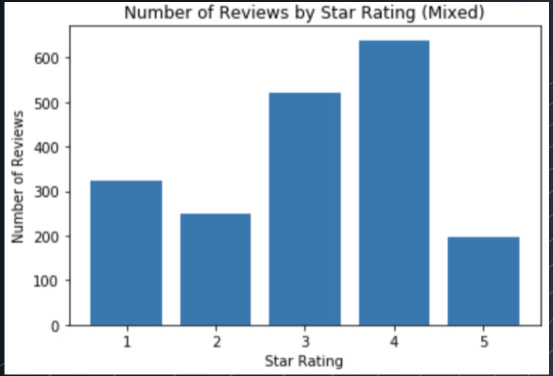
Positive-Negative Inflection Point occurs around 3.

# Rating Distribution by Sentiment


Number of Reviews by Star Rating (Positive)


Number of Reviews by Star Rating (Neutral)

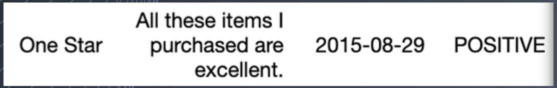

Number of Reviews by Star Rating (Negative)


Number of Reviews by Star Rating (Mixed)

**Observation:**

**Unusual** occurrences (high star negative reviews / low star positive reviews) likely due to human error upon closer inspection.

Example:



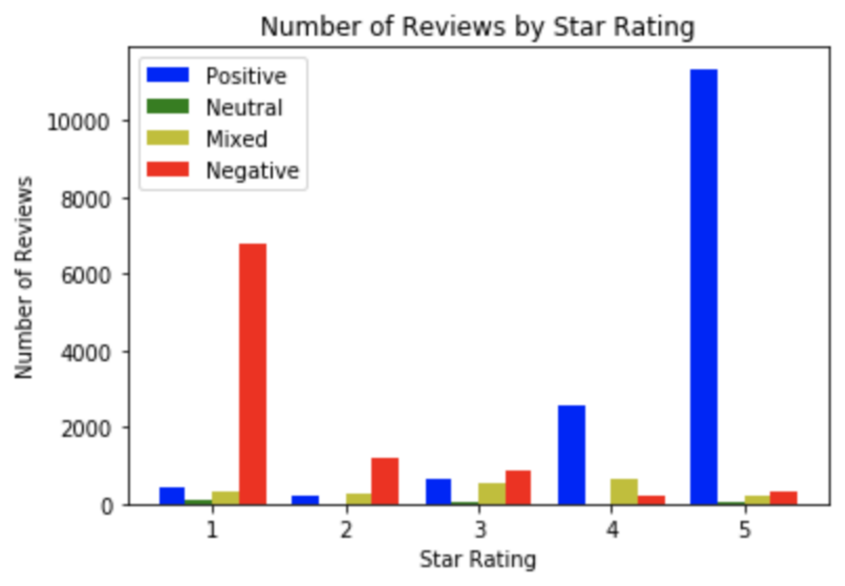
| Rating | Comment | Date | Sentiment |
|---|---|---|---|
| One Star | All these items I purchased are excellent. | 2015-08-29 | POSITIVE |

# 5.

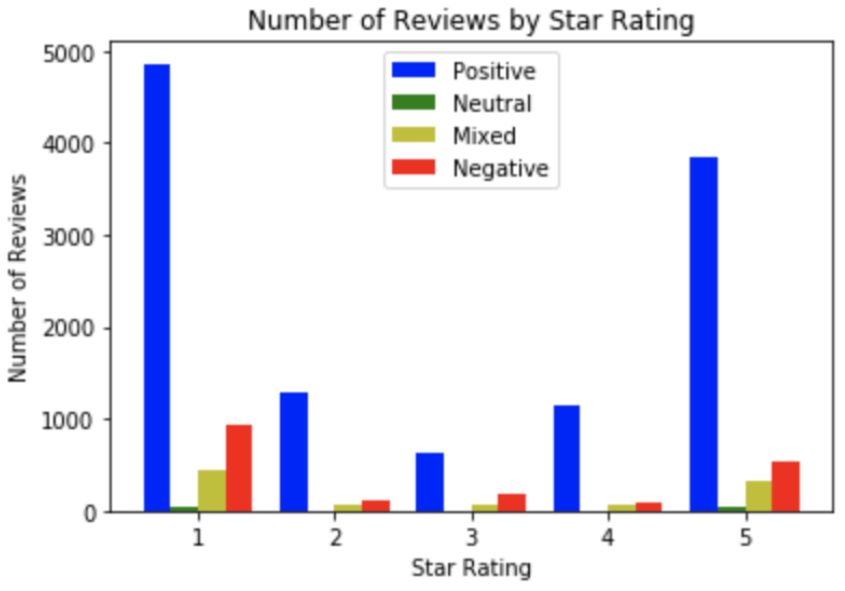
# Cross-category Sentiment Analysis

# Rating Distribution Comparison:

### Major Appliances:



### Video Games:



# Rating Comparison

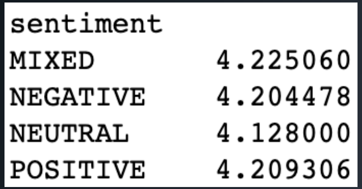
**Average Star Rating Comparison:**

**3.8853 - Major Appliances**

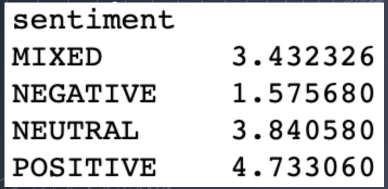
**4.2039 - Video Games**

**Star Rating becomes unreliable for products that change over time:**

**Major Appliances:**



```
sentiment
MIXED        4.225060
NEGATIVE     4.204478
NEUTRAL      4.128000
POSITIVE     4.209306
```



```
sentiment
MIXED        3.432326
NEGATIVE     1.575680
NEUTRAL      3.840580
POSITIVE     4.733060
```

Rating Expectations greatly vary depending on product category

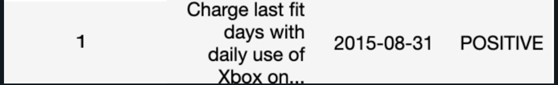
- This could be due to various factors, such as:
  - availability of choice (20,000 racing games vs 5 washing machines)
  - ease of review (finding scratches on product vs visual glitch in game)

# Adjustable Ratings:

## Digital Release Items

**Seller follows up** following bad rating with remedy action, causing review/rating to alter. User may forget to change either/both.

Example:



| Rating | Comment | Date | Sentiment |
|---|---|---|---|
| 1 | Charge last fit days with daily use of Xbox on... | 2015-08-31 | POSITIVE |

Alternatively, some products (e.g. digital releases) can change over time.

Result: Unreliable ratings that require further analysis

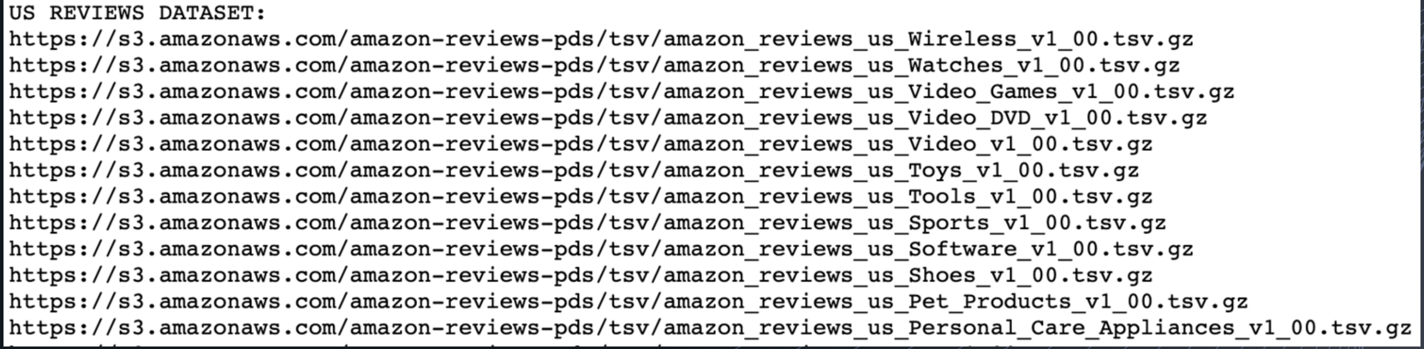
# Further Analysis

Cross-Category comparison will become very interesting when done at scale

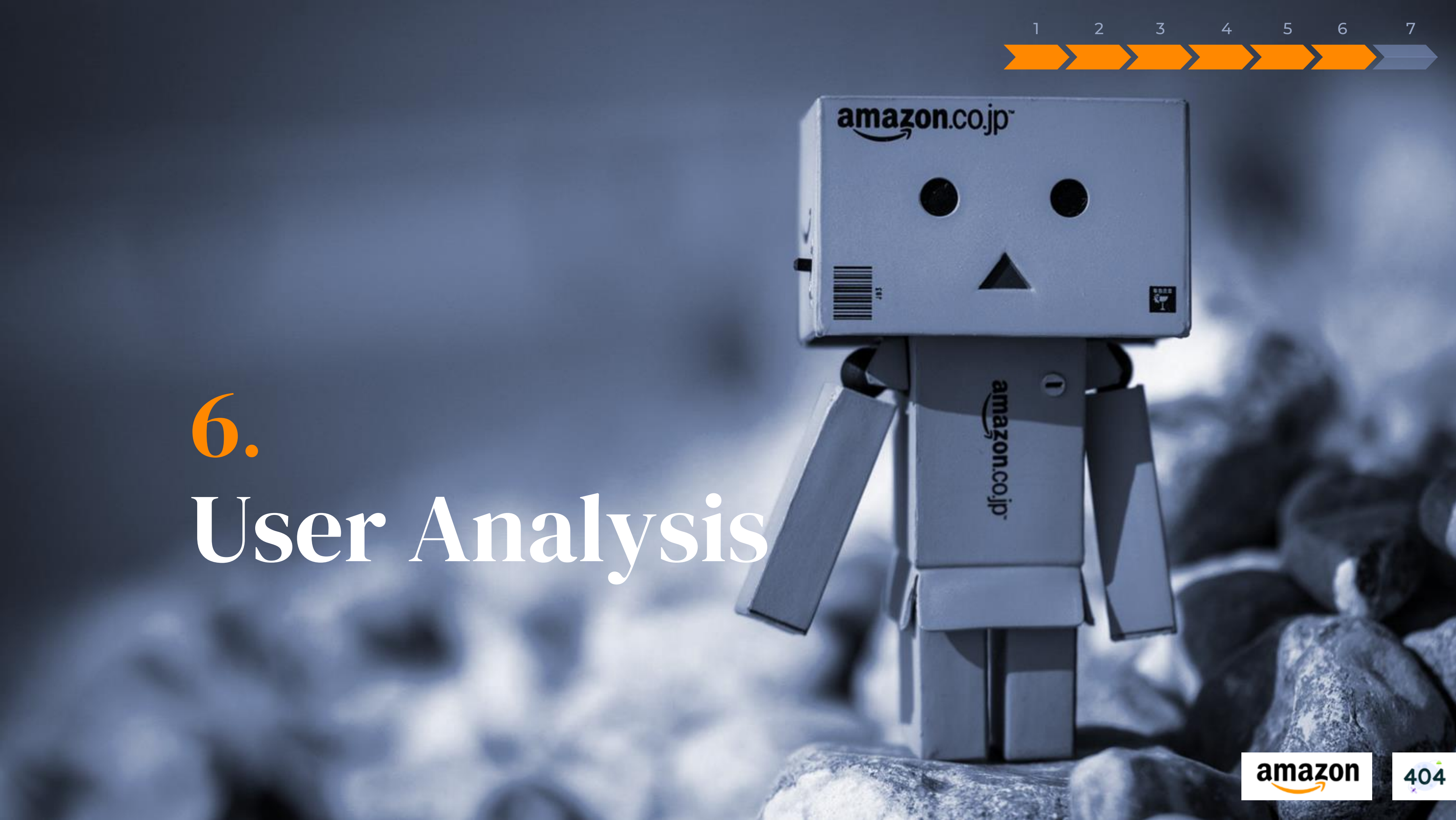
- Reviews can be adjusted based per-category to better reflect reality
- Analyse review and sentiments by customer profiling (Jie)

We are constrained by our budget on this project:

Every ~10,000 lines costed roughly 6-7 USD for Amazon Comprehend to process



```
US REVIEWS DATASET:
https://s3.amazonaws.com/amazon-reviews-pds/tsv/amazon_reviews_us_Wireless_v1_00.tsv.gz
https://s3.amazonaws.com/amazon-reviews-pds/tsv/amazon_reviews_us_Watches_v1_00.tsv.gz
https://s3.amazonaws.com/amazon-reviews-pds/tsv/amazon_reviews_us_Video_Games_v1_00.tsv.gz
https://s3.amazonaws.com/amazon-reviews-pds/tsv/amazon_reviews_us_Video_DVD_v1_00.tsv.gz
https://s3.amazonaws.com/amazon-reviews-pds/tsv/amazon_reviews_us_Video_v1_00.tsv.gz
https://s3.amazonaws.com/amazon-reviews-pds/tsv/amazon_reviews_us_Toys_v1_00.tsv.gz
https://s3.amazonaws.com/amazon-reviews-pds/tsv/amazon_reviews_us_Tools_v1_00.tsv.gz
https://s3.amazonaws.com/amazon-reviews-pds/tsv/amazon_reviews_us_Sports_v1_00.tsv.gz
https://s3.amazonaws.com/amazon-reviews-pds/tsv/amazon_reviews_us_Software_v1_00.tsv.gz
https://s3.amazonaws.com/amazon-reviews-pds/tsv/amazon_reviews_us_Shoes_v1_00.tsv.gz
https://s3.amazonaws.com/amazon-reviews-pds/tsv/amazon_reviews_us_Pet_Products_v1_00.tsv.gz
https://s3.amazonaws.com/amazon-reviews-pds/tsv/amazon_reviews_us_Personal_Care_Appliances_v1_00.tsv.gz
```



# 6.
# User Analysis

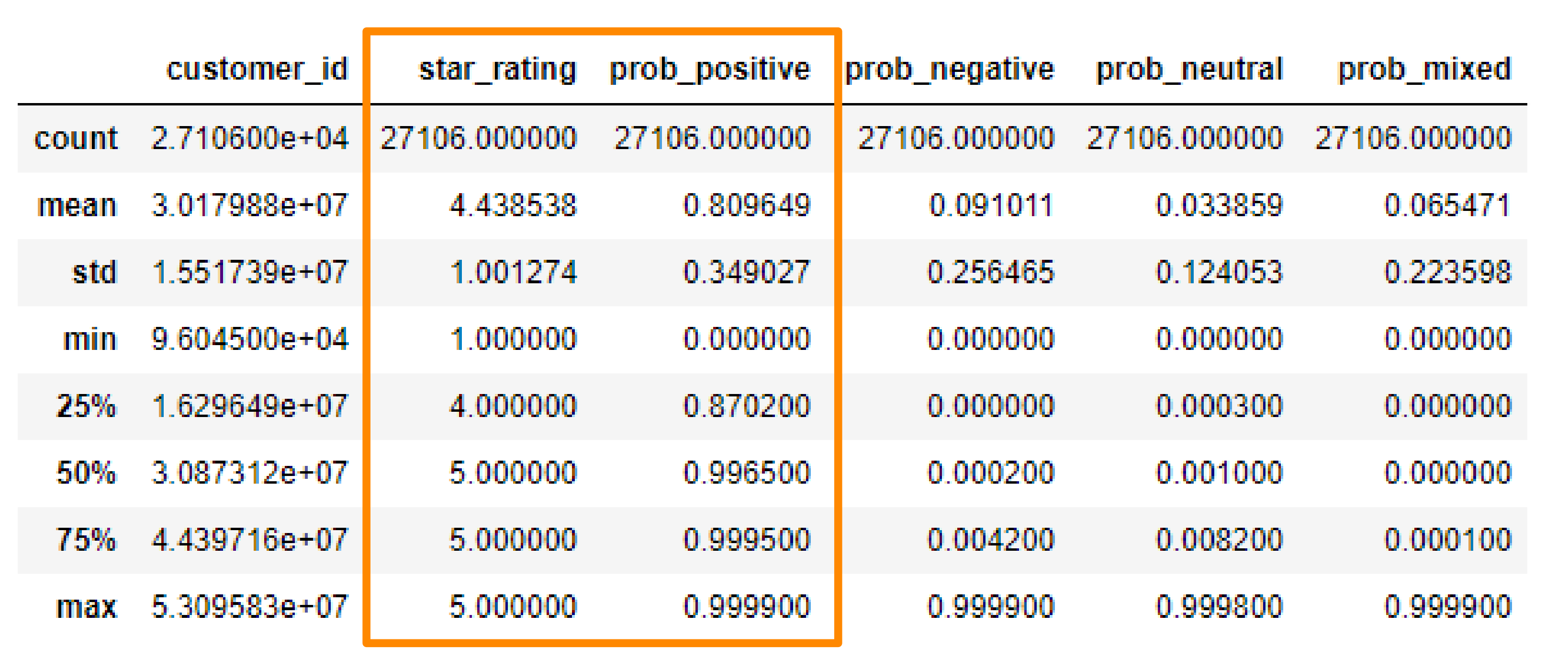
# Data Used

**Index Page:** https://s3.amazonaws.com/amazon-reviews-pds/tsv/index.txt

**Datasets:**

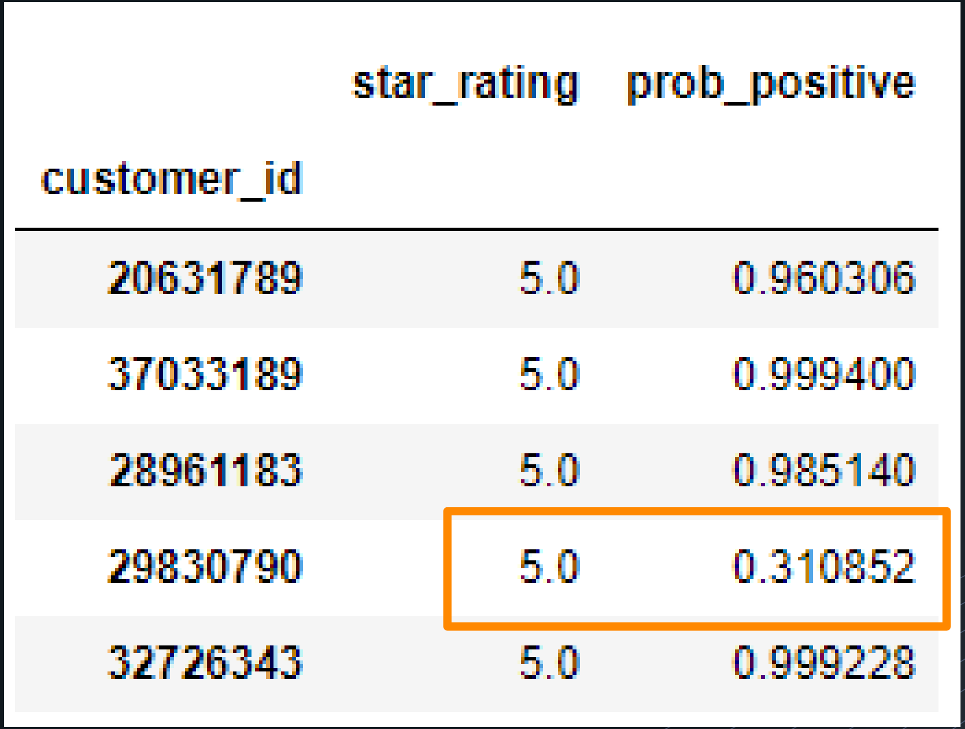
- 46 US review datasets
- 5 multilingual review datasets
- 4 US review datasets used
  - Major Appliances
  - Furniture
  - Watches
  - Musical instruments
- 2,752,353 rows initially
- 27,106 rows after (customer id with more than 30 reviews)
- 569 unique customers

# Overview



| | customer_id | star_rating | prob_positive | prob_negative | prob_neutral | prob_mixed |
|---|---|---|---|---|---|---|
| count | 2.710600e+04 | 27106.000000 | 27106.000000 | 27106.000000 | 27106.000000 | 27106.000000 |
| mean | 3.017988e+07 | 4.438538 | 0.809649 | 0.091011 | 0.033859 | 0.065471 |
| std | 1.551739e+07 | 1.001274 | 0.349027 | 0.256465 | 0.124053 | 0.223598 |
| min | 9.604500e+04 | 1.000000 | 0.000000 | 0.000000 | 0.000000 | 0.000000 |
| 25% | 1.629649e+07 | 4.000000 | 0.870200 | 0.000000 | 0.000300 | 0.000000 |
| 50% | 3.087312e+07 | 5.000000 | 0.996500 | 0.000200 | 0.001000 | 0.000000 |
| 75% | 4.439716e+07 | 5.000000 | 0.999500 | 0.004200 | 0.008200 | 0.000100 |
| max | 5.309583e+07 | 5.000000 | 0.999900 | 0.999900 | 0.999800 | 0.999900 |

# Highest Star-Rating Givers



| | star_rating | prob_positive |
|---|---|---|
| customer_id | | |
| 20631789 | 5.0 | 0.960306 |
| 37033189 | 5.0 | 0.999400 |
| 28961183 | 5.0 | 0.985140 |
| 29830790 | 5.0 | 0.310852 |
| 32726343 | 5.0 | 0.999228 |

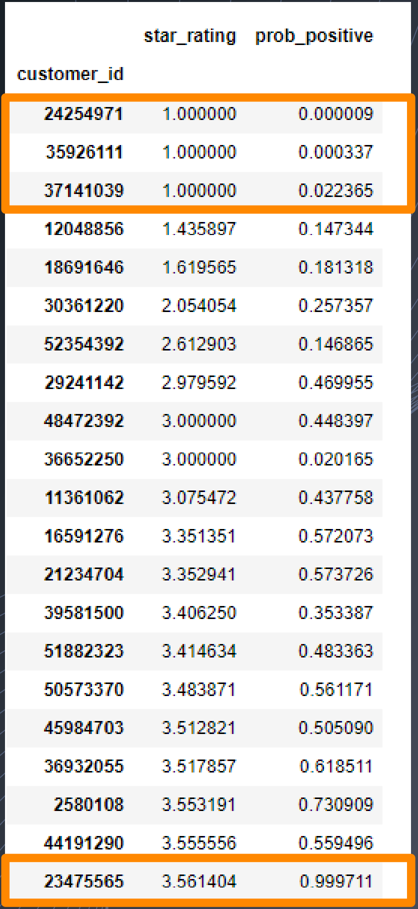
Among all 569 customers with more than 30 reviews:

- Highest star-rating is 5 stars
- 51 customers always give 5 stars
- Most 5-star raters have also very high average positive sentiment probability
- Some customers have low positive sentiment probability
- They rate 5-star even when they are not satisfied

# Lowest Star-Rating Givers

Among all 569 customers with more than 30 reviews:

- Lowest star-rating is 1 star
- 3 customers always give 1 star
- Most low-star raters have also low average positive sentiment probability
- Some customers have very high positive sentiment probability
- They rate low even when they are satisfied OR
- They are being polite in the review section



| | star_rating | prob_positive |
|---|---|---|
| customer_id | | |
| 24254971 | 1.000000 | 0.000009 |
| 35926111 | 1.000000 | 0.000337 |
| 37141039 | 1.000000 | 0.022365 |
| 12048856 | 1.435897 | 0.147344 |
| 18691646 | 1.619565 | 0.181318 |
| 30361220 | 2.054054 | 0.257357 |
| 52354392 | 2.612903 | 0.146865 |
| 29241142 | 2.979592 | 0.469955 |
| 48472392 | 3.000000 | 0.448397 |
| 36652250 | 3.000000 | 0.020165 |
| 11361062 | 3.075472 | 0.437758 |
| 16591276 | 3.351351 | 0.572073 |
| 21234704 | 3.352941 | 0.573726 |
| 39581500 | 3.406250 | 0.353387 |
| 51882323 | 3.414634 | 0.483363 |
| 50573370 | 3.483871 | 0.561171 |
| 45984703 | 3.512821 | 0.505090 |
| 36932055 | 3.517857 | 0.618511 |
| 2580108 | 3.553191 | 0.730909 |
| 44191290 | 3.555556 | 0.559496 |
| 23475565 | 3.561404 | 0.999711 |

# Toxic Star-Rating Givers

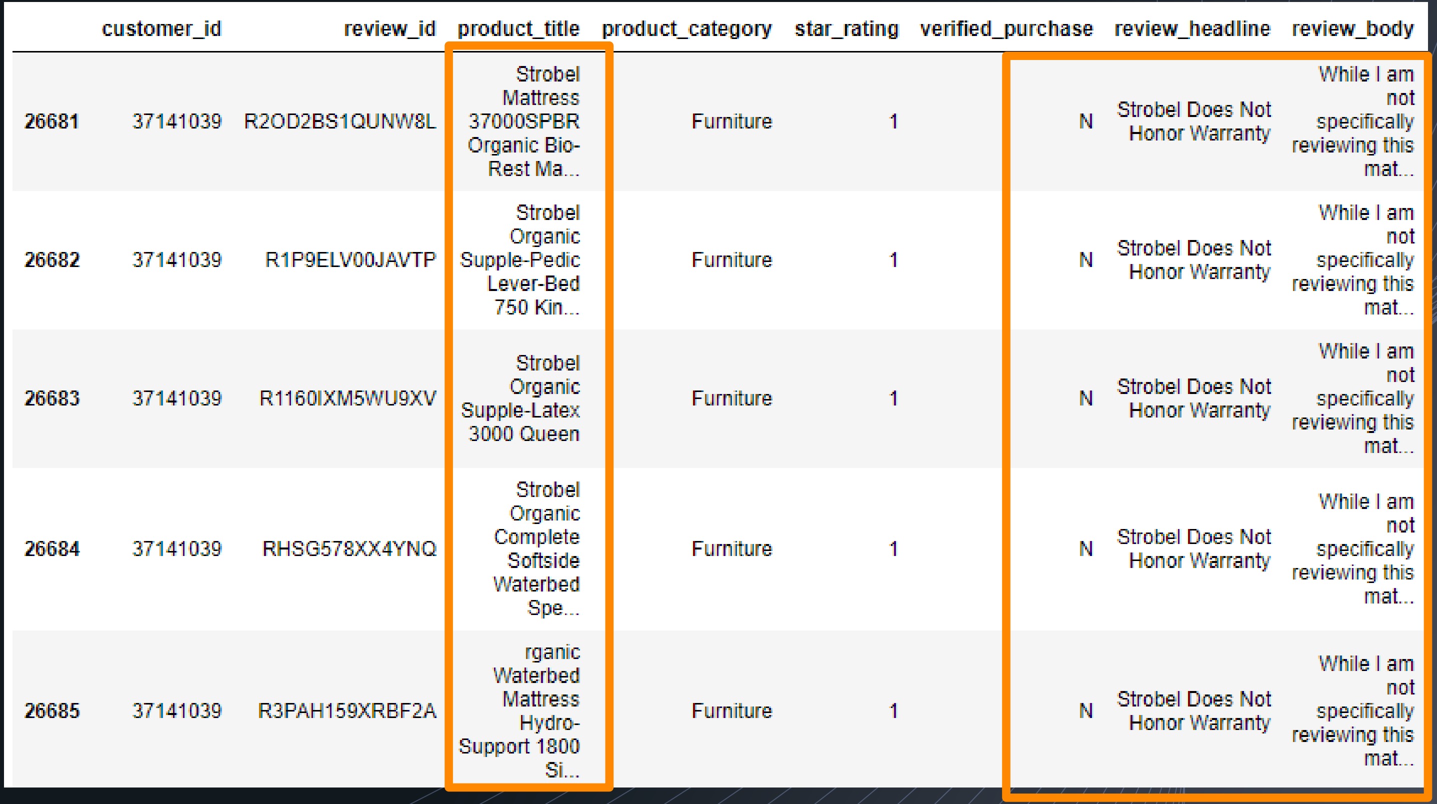
Among 3 customers with all 1-star reviews:

- Customer 24254971 only reviewed watches
- Customer 35926111 only reviewed watches
- Customer 37141039 only reviewed furniture, only mattress in particular
- Their reviews are all without Verified Purchase
- It could be that they are very picky about watches and mattress

OR

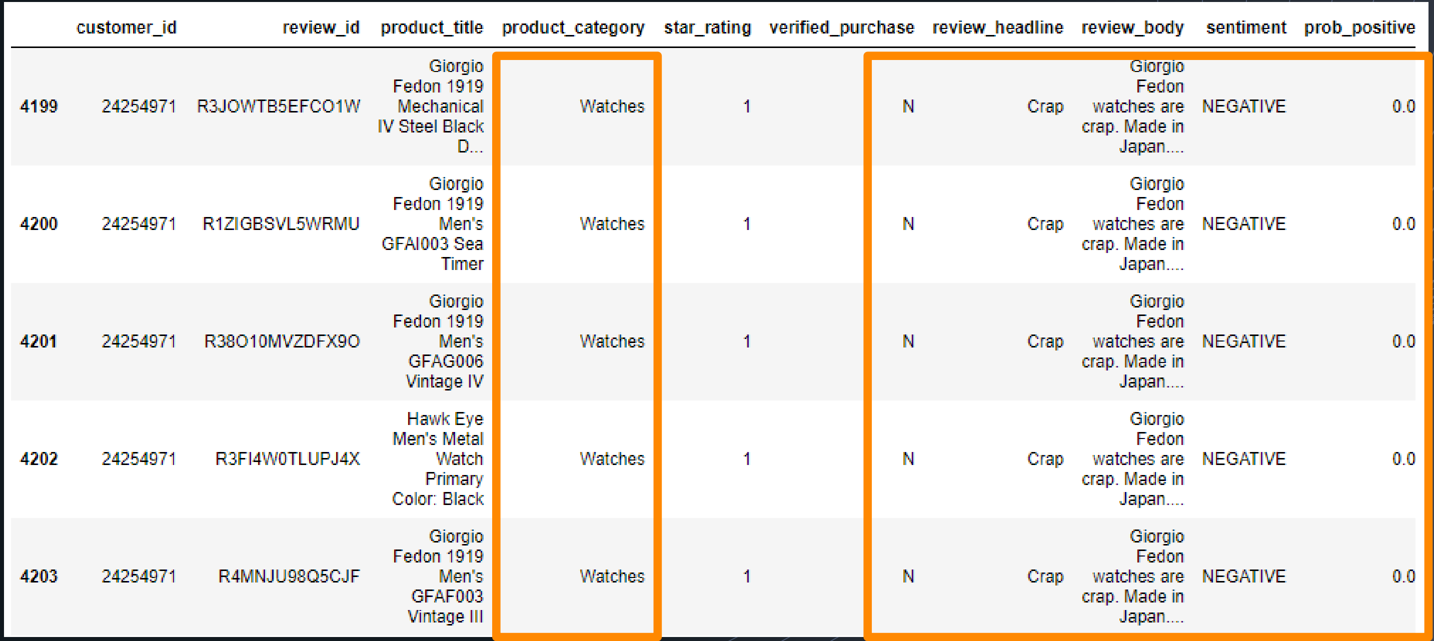
- They are giving 1-star on purpose

# Toxic Star-Rating Givers



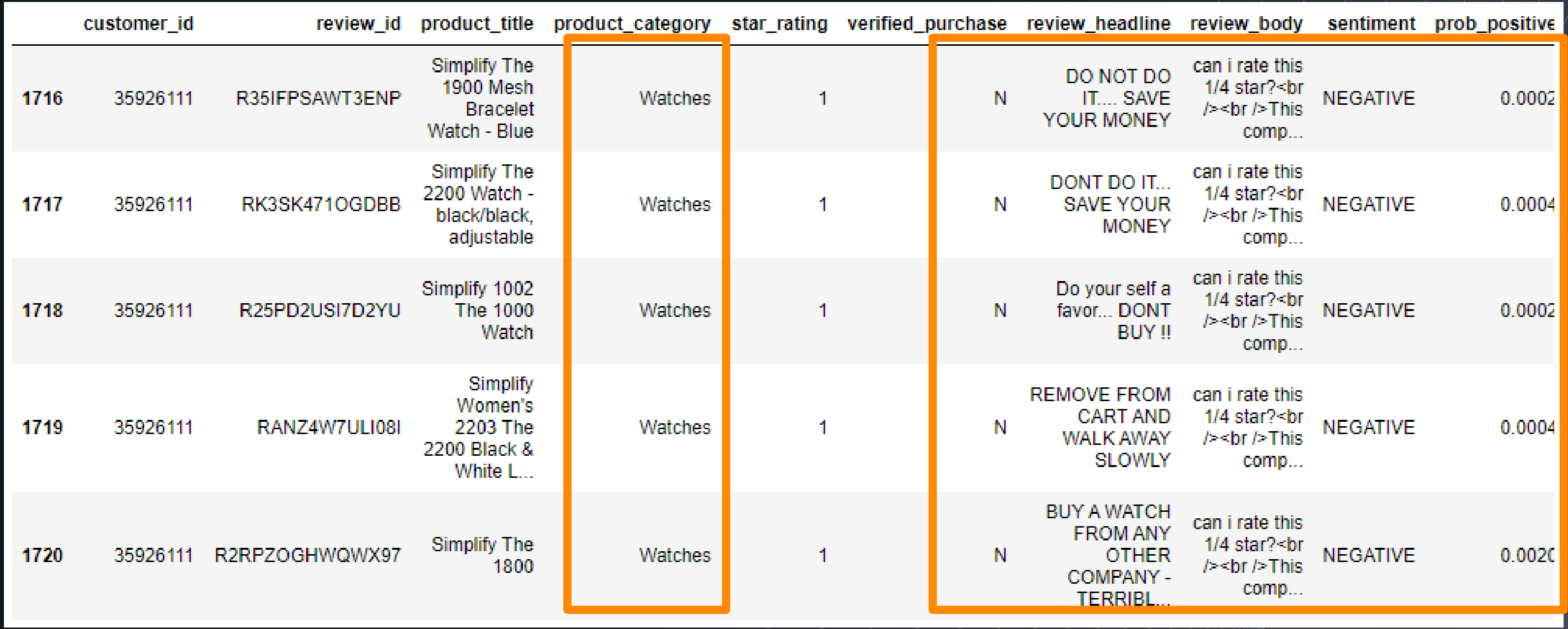
| | customer_id | review_id | product_title | product_category | star_rating | verified_purchase | review_headline | review_body |
|---|---|---|---|---|---|---|---|---|
| **26681** | 37141039 | R2OD2BS1QUNW8L | Strobel Mattress 37000SPBR Organic Bio-Rest Ma... | Furniture | 1 | N | Strobel Does Not Honor Warranty | While I am not specifically reviewing this mat... |
| **26682** | 37141039 | R1P9ELV00JAVTP | Strobel Organic Supple-Pedic Lever-Bed 750 Kin... | Furniture | 1 | N | Strobel Does Not Honor Warranty | While I am not specifically reviewing this mat... |
| **26683** | 37141039 | R1160IXM5WU9XV | Strobel Organic Supple-Latex 3000 Queen | Furniture | 1 | N | Strobel Does Not Honor Warranty | While I am not specifically reviewing this mat... |
| **26684** | 37141039 | RHSG578XX4YNQ | Strobel Organic Complete Softside Waterbed Spe... | Furniture | 1 | N | Strobel Does Not Honor Warranty | While I am not specifically reviewing this mat... |
| **26685** | 37141039 | R3PAH159XRBF2A | rganic Waterbed Mattress Hydro-Support 1800 Si... | Furniture | 1 | N | Strobel Does Not Honor Warranty | While I am not specifically reviewing this mat... |

# Toxic Star-Rating Givers



| | customer_id | review_id | product_title | product_category | star_rating | verified_purchase | review_headline | review_body | sentiment | prob_positive |
|---|---|---|---|---|---|---|---|---|---|---|
| **4199** | 24254971 | R3JOWTB5EFCO1W | Giorgio Fedon 1919 Mechanical IV Steel Black D... | Watches | 1 | N | Crap | Giorgio Fedon watches are crap. Made in Japan.... | NEGATIVE | 0.0 |
| **4200** | 24254971 | R1ZIGBSVL5WRMU | Giorgio Fedon 1919 Men's GFAI003 Sea Timer | Watches | 1 | N | Crap | Giorgio Fedon watches are crap. Made in Japan.... | NEGATIVE | 0.0 |
| **4201** | 24254971 | R38O10MVZDFX9O | Giorgio Fedon 1919 Men's GFAG006 Vintage IV | Watches | 1 | N | Crap | Giorgio Fedon watches are crap. Made in Japan.... | NEGATIVE | 0.0 |
| **4202** | 24254971 | R3FI4W0TLUPJ4X | Hawk Eye Men's Metal Watch Primary Color: Black | Watches | 1 | N | Crap | Giorgio Fedon watches are crap. Made in Japan.... | NEGATIVE | 0.0 |
| **4203** | 24254971 | R4MNJU98Q5CJF | Giorgio Fedon 1919 Men's GFAF003 Vintage III | Watches | 1 | N | Crap | Giorgio Fedon watches are crap. Made in Japan.... | NEGATIVE | 0.0 |

# Toxic Star-Rating Givers



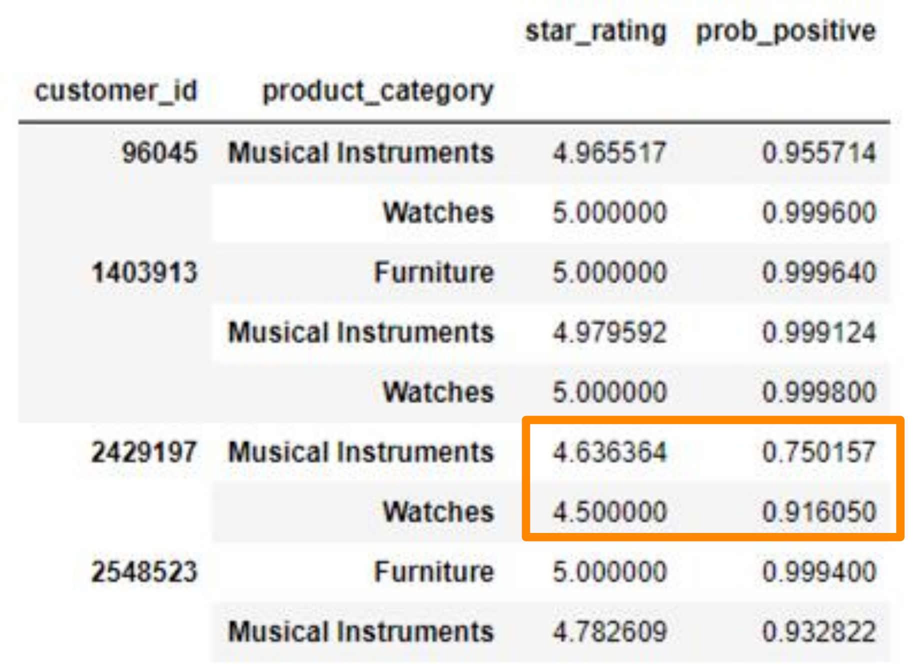
| | customer_id | review_id | product_title | product_category | star_rating | verified_purchase | review_headline | review_body | sentiment | prob_positive |
|---|---|---|---|---|---|---|---|---|---|---|
| **1716** | 35926111 | R35IFPSAWT3ENP | Simplify The 1900 Mesh Bracelet Watch - Blue | Watches | 1 | N | DO NOT DO IT.... SAVE YOUR MONEY | can i rate this 1/4 star?<br /><br />This comp... | NEGATIVE | 0.0002 |
| **1717** | 35926111 | RK3SK471OGDBB | Simplify The 2200 Watch - black/black, adjustable | Watches | 1 | N | DONT DO IT... SAVE YOUR MONEY | can i rate this 1/4 star?<br /><br />This comp... | NEGATIVE | 0.0004 |
| **1718** | 35926111 | R25PD2USI7D2YU | Simplify 1002 The 1000 Watch | Watches | 1 | N | Do your self a favor... DONT BUY !! | can i rate this 1/4 star?<br /><br />This comp... | NEGATIVE | 0.0002 |
| **1719** | 35926111 | RANZ4W7ULI08I | Simplify Women's 2203 The 2200 Black & White L... | Watches | 1 | N | REMOVE FROM CART AND WALK AWAY SLOWLY | can i rate this 1/4 star?<br /><br />This comp... | NEGATIVE | 0.0004 |
| **1720** | 35926111 | R2RPZOGHWQWX97 | Simplify The 1800 | Watches | 1 | N | BUY A WATCH FROM ANY OTHER COMPANY - TERRIBL... | can i rate this 1/4 star?<br /><br />This comp... | NEGATIVE | 0.0020 |

# Toxic Star-Rating Givers

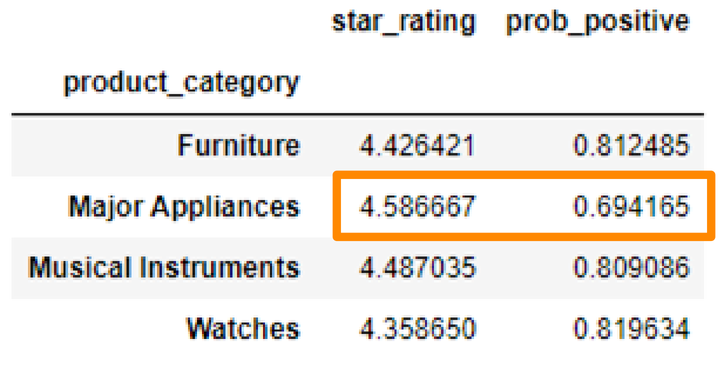
can i rate this 1/4 star?

This company is in the business of discounting their (absolute BS) $200 watch down so you think you getting a good deal. In actuality they sell you a terribly constructed piss poor product. Im pretty sure they are trying to knock off a chinese knock off of a american made watch. THATS TWO KNOCK OFFS. Watch broke after 1 month. 1 e'Fing MONTH!!?!?!? Absolutely fell apart. I emailed customer service and they told me i could buy a replacement and they wouldn't do anything for me. Sorry but when i buy products i expect them to last AT LEAST more than a month of gentle wear (office job). Terrible company, terrible watch.<br /><br />Honestly you'll have better luck putting a quarter in the machine at walmart and getting the little pill bottle watch. At least that one will last.

# Customer Reviews in Multiple Categories



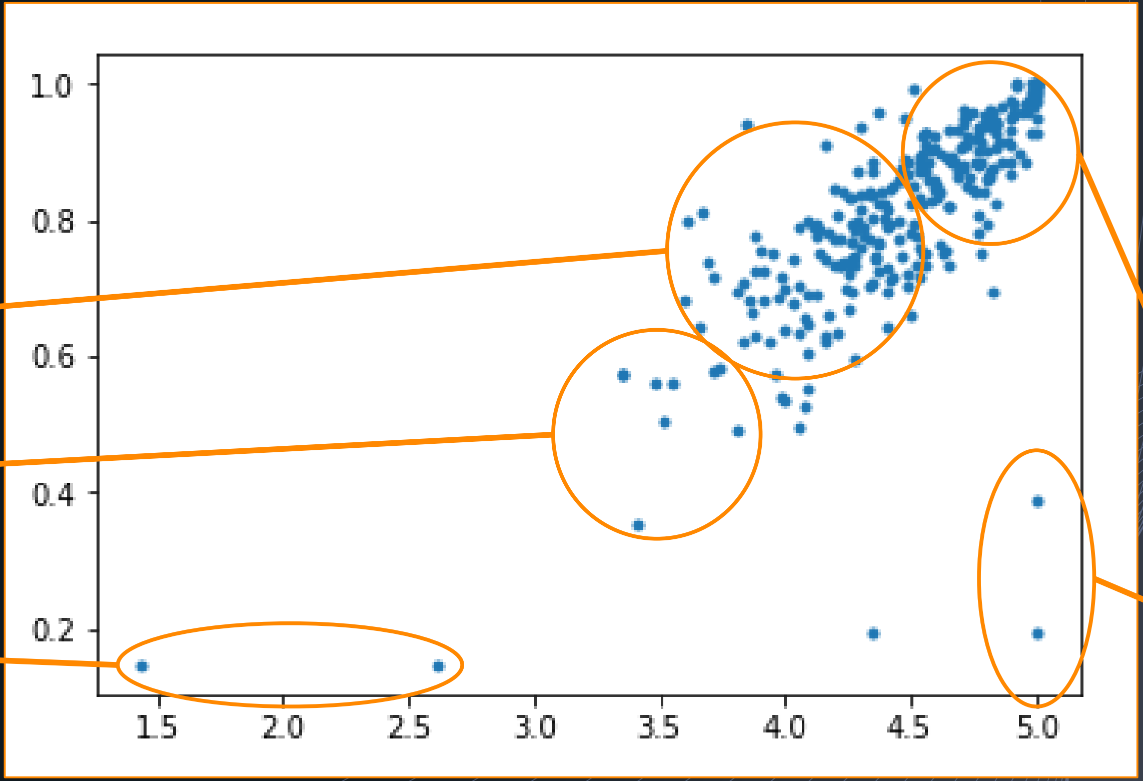
| | | star_rating | prob_positive |
|---|---|---|---|
| customer_id | product_category | | |
| 96045 | Musical Instruments | 4.965517 | 0.955714 |
| | Watches | 5.000000 | 0.999600 |
| 1403913 | Furniture | 5.000000 | 0.999640 |
| | Musical Instruments | 4.979592 | 0.999124 |
| | Watches | 5.000000 | 0.999800 |
| 2429197 | Musical Instruments | 4.636364 | 0.750157 |
| | Watches | 4.500000 | 0.916050 |
| 2548523 | Furniture | 5.000000 | 0.999400 |
| | Musical Instruments | 4.782609 | 0.932822 |



| | star_rating | prob_positive |
|---|---|---|
| product_category | | |
| Furniture | 4.426421 | 0.812485 |
| Major Appliances | 4.586667 | 0.694165 |
| Musical Instruments | 4.487035 | 0.809086 |
| Watches | 4.358650 | 0.819634 |

- Include only customers with more than 30 reviews in 2 or more categories
- 285 customers
- 14,442 rows

# Star-Sentiment Plot



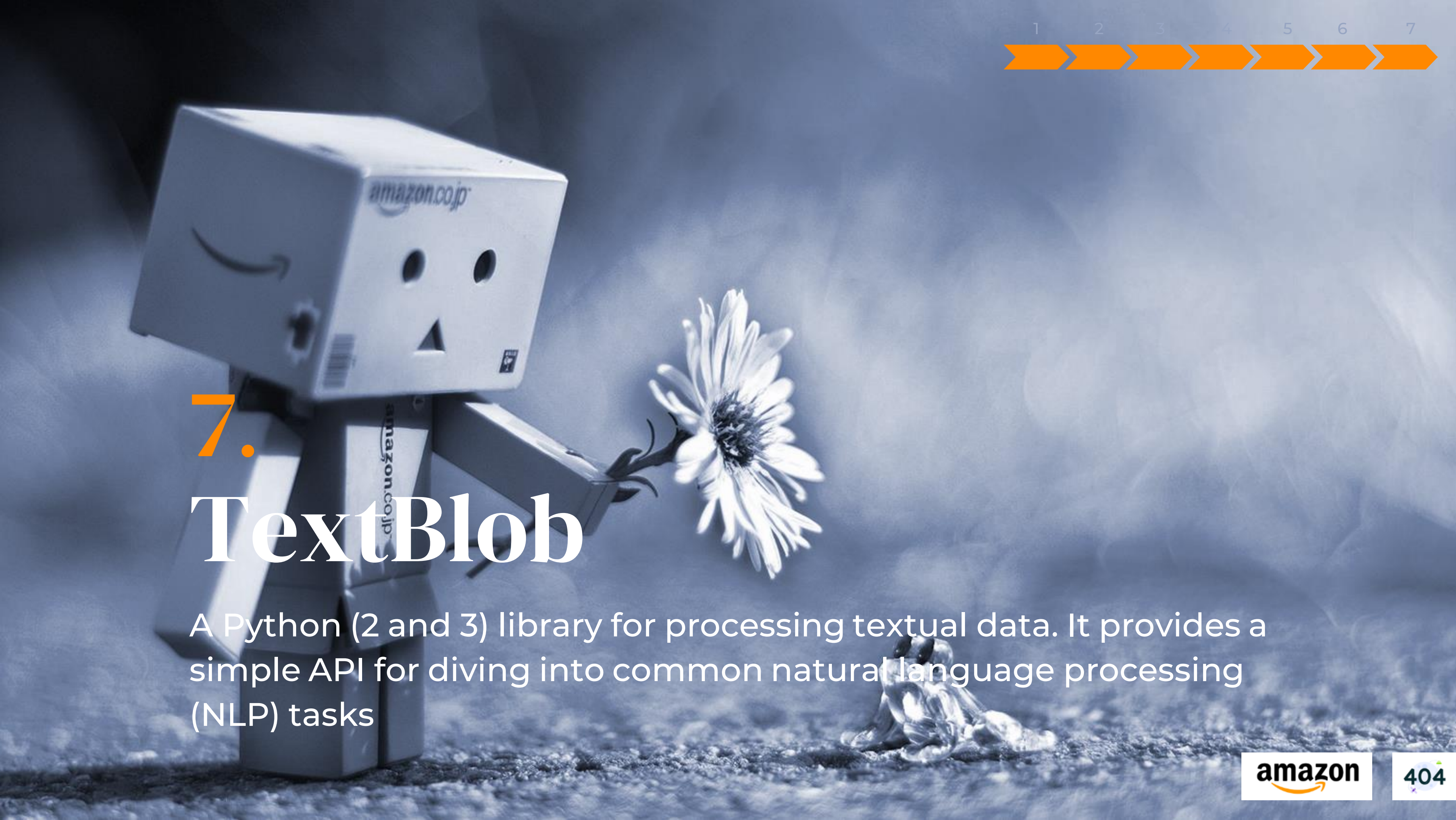
Average

Hard-to-Satisfied

Grumpy

Easily - Satisfied

5-Star Giver



# 7.

# TextBlob

A Python (2 and 3) library for processing textual data. It provides a simple API for diving into common natural language processing (NLP) tasks

# More tools for NLP

- NLTK
- Spacy
- Stanford Core NLP
- TextBlob

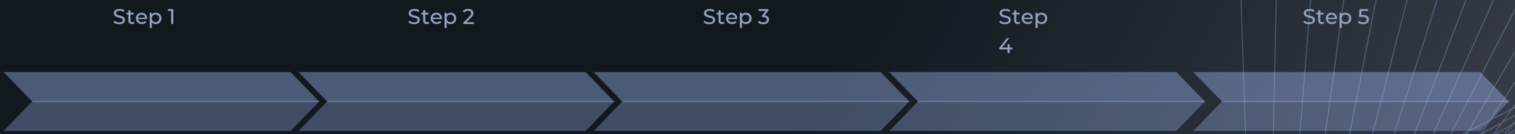
- TextBlob:
  - Easier to use because it has nicer user interface
  - Documentation is thoroughly explained
  - Built on the shoulders of NLTK

# Features of TextBlob

1. Noun phrase extraction
2. Part-of-speech tagging
3. Sentiment analysis
4. Classification (Naive Bayes, Decision Tree)
5. Language translation and detection powered by Google Translate
6. Tokenization (splitting text into words and sentences)
7. Word and phrase frequencies
8. Parsing
9. n-grams
10. Word inflection (pluralization and singularization) and lemmatization
11. Spelling correction
12. Add new models or languages through extensions
13. WordNet integration

Use sentiment analysis from TextBlob in Pyspark to analysis Amazon review data and **compare** its results with Amazon Comprehend

# Sentiment Analysis Pipeline



### Data Collection

Amazon Review Data

### Text Preparation

- Fix abbreviation
- Remove irrelevant features

### Sentiment Detection

- PatternAnalyzer
- polarity, subjectivity

### Sentiment Classification

- NaiveBayesAnalyzer
- Based on Score
  - >0.1 Positive
  - [-0.1, 0.1] Neutral
  - <-0.1 Negative

### Compare Results

# Compare the count of each category

| | Amazon Comprehend | TextBlob |
|---|---|---|
| Positive | 6696 | 6911 |
| Neutral | 69 | 2331 |
| Negative | 2306 | 758 |
| Mixed | 749 | |

# Take a closer look at sentiment detection:

| Review Text | Amazon Comprehend | TextBlob |
|---|---|---|
| Love my refrigerator! ! Keeps everything cold..will recommend! | Positive | Neutral |
| AS advertised | Positive | Neutral |
| It's not worth 22 dollars, I've heard it became of some value just not that high. | Negative | Neutral |
| Did the job but didn't match the original gray wheels. | Mixed | Positive |
| Cheap knock-off. Don’t waste your time | Negative | Positive |

# Comparison Summary

- Amazon Comprehend has a better sentiment detection accuracy

  Possible reason:

  1. It’s based on Machine Learning while TextBlob(PatternAnalyzer) is based on dictionary)
  2. Amazon comprehend has a custom set of entities or text classification models that are tailored uniquely to text data

- Amazon Comprehend has the level “Mixed”, which better classifies reviews

- TextBlob is **free**

# Next Steps

- Current Constraints:
  - AWS Educate only supports t2.medium for Jupyter Notebook instances
  - AWS Comprehend costs about $6 per 10000 reviews
- Future Directions:
  - Analyze multilingual datasets
  - Analyze all 46 US datasets
  - Track customer review patterns across categories for better segmentation
  - Analyse whether review patterns differs by barriers (e.g. platform, region, etc.)
  - Analyse whether review patterns have shifted over time

# Thank You

Q&A